Statistics and Probability Letters 144 (2019) 57–62

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Functional analysis of spatial aggregation regions of Jeffrey pine beetle-attack within the Lake Tahoe Basin

Ekaterina Smirnova [a,*], Omid Khormali [a], Joel M. Egan [b]

[a] Department of Mathematical Sciences, University of Montana, 32 Campus dr., Missoula, MT, USA
[b] Forest Health Protection, USDA Forest Service, 26 Fort Missoula Road, Missoula, MT, USA

ARTICLE INFO

Article history:
Available online 6 July 2018

Keywords:
Spatial aggregation
Functional data analysis
Function-on-function regression
Complex domain modeling
Bark beetles
Dendroctonus jeffreyi

ABSTRACT

We introduce modeling approaches for quantifying the spatiotemporal location and characteristics of clusters with application to the 1993 Jeffrey pine beetle (JPB) forest epidemic attack.

1. Introduction

Modeling the location, characteristics, and dynamics of clusters that occur during ecological outbreaks is an important topic in environmental sciences. Motivated by the 1993 Jeffrey pine beetle (JPB) forest epidemic attack in the Lake Tahoe Basin, we propose methods that describe the location, shape, and characteristics of spatial clusters. Our purpose is to introduce a novel functional representation approach to describe the complex shape and characteristics of spatial clusters. We start with n previously defined clusters, and for each cluster, we first separate the domain into G non-overlapping cones located at the cluster center, and then describe the contour as a distance function at the angle θ_g, $g = 1, \ldots, G$, that defines each cone. Additional information about the cluster (e.g., number of affected trees) can be collected and represented as a function of the cone-specific angle. This novel approach of clusters size and shape representation as functional data allows modeling the cluster characteristics (e.g., shape, size, location), and covariates (e.g., forest attributes) in each direction.

In recent years, the study of irregularly shaped clusters and development of spatial clusters detection without using fixed geometries has received considerable attention (Duczmal et al., 2006; Yiannakoulias et al., 2010; Duarte et al., 2010; Moreira et al., 2015; Lawson, 2010). Methods were inspired by and applied to the study of geographical regions that define disease or ecological clusters. Applications of irregularly shaped cluster detection methods span a variety of disciplines including breast cancer mortality data from northeastern United States (Duczmal et al., 2006; Duarte et al., 2010), Schistosoma mansoni and hookworm infection intensities in Kenya (Yiannakoulias et al., 2010), Chagas' disease clusters in Brazil (Moreira et al., 2015), and ecology, health and environmental genomics problems (Lawson, 2010). The study of spatial clusters formation processes has been shown to be important in the context of industrial agglomerations (Kerr and Kominers, 2015; Mori and Smith, 2013). In particular, Kerr and Kominers (2015) model the spatial clusters formation in the context of industrial firms interactions based on physical proximity and similarity between firms. Clusters are built from overlapping regions of firms microinteractions and the model provides connection between observable cluster shapes and the underlying agglomerative forces that cause them. A probabilistic model of multiple clusters for the spatial analysis of industrial agglomeration patterns

* Corresponding author.
E-mail address: ekaterina.smirnova@mso.umt.edu (E. Smirnova).

https://doi.org/10.1016/j.spl.2018.07.001
0167-7152/

is introduced in Mori and Smith (2013) and an overview of distance based methods for studying spatial concentration is provided in Marcon and Puech (2017). Methodologically, these approaches are concerned with the detection of spatial clusters. In contrast, our clusters have been identified manually by scientists with detailed knowledge of the JPB infestation and we are interested in modeling their characteristics and cross-sectional and longitudinal association with covariates. Our novel cluster representation approach, provides a new inferential and computational platform that was previously unavailable and makes a direct connection to functional data analysis. This connection allows us to use the inferential and computational machinery already existent in functional data analysis to conduct inference for complex-structured spatially clustered data.

An alternative approach to inference over an irregular domain is *spatial smoothing*. The standard model for spatial spline smoothing function over irregularly shaped spatial domain $\Omega_i \subseteq \mathcal{R}$ is

$$z_{ij} = f(\boldsymbol{p}_{ij}) + \epsilon_{ij}, \quad j = 1, \ldots, m_i, \tag{1}$$

where the data are $\{(\boldsymbol{p}_{i1}, z_{i1}), \ldots, (\boldsymbol{p}_{im_i}, z_{im_i})\} \subset \Omega_i \times \mathcal{R}$, $\epsilon_{ij} \sim (0, \sigma_i^2)$ are independent errors, and $f(\cdot)$ is a sufficiently smooth function from $\Omega_i \to \mathcal{R}$ (Ramsay, 2002). In this model, the function value z_{ij} is observed at points $\boldsymbol{p}_{ij} = (x_{ij}, y_{ij}), j = 1, \ldots, m_i$ within the irregular domain Ω_i. In the context of the spatial clustering problem, each cluster is the irregularly shaped domain $\Omega_i, i = 1, \ldots, n$ and the response z_{ij} is for points $\boldsymbol{p}_{ij}$ within cluster i. The spatial smoothing literature considers the models within one domain only, that is $i = 1$. The standard spatial smoothing model (1) was extended in Sangalli et al. (2013) to include covariates associated with each observation z_{ij} and further extended by Bhattacharjee et al. (2017) to estimate the functional surface of a regression coefficient that varies over a domain that includes discontinuities in the spatial surface. Our approach is substantially different from the smoothing approach, because our data are very precisely observed. The lack of smoothing either in the edges of the clusters or in the interior values is a representation of the actual data. Instead, we define precisely the shape of the cluster and its associated characteristics and impose smoothing on the functional parameters that describe the association between cluster characteristics and covariates.

This paper is organized as follows. We describe motivating data example in Section 2. In Section 3, we introduce methods for describing the size and shape of clusters over irregular domains. We present our method's application to the analysis of spatial aggregation regions of Jeffrey pine beetle-attack in Section 4. Discussion and concluding remarks follow in Section 5.

2. Data description

We consider the census data collected on 10,721 trees in the Lake Tahoe Basin during the 1991–1996 Jeffrey pine beetle (*Dendroctonus jeffreyi* Hopkins) (JPB) outbreak. This outbreak caused *caused severe tree mortality* during protracted drought conditions (Egan et al., 2016). Severe mortality occurrences are of interest as they conflict with resource management objectives related to wildlife habitat conservation, fuel structure and loading regulation, carbon sequestration, timber production, or recreation. JPB outbreaks have progressive spatial clustering that occurs across two epidemiological phases that are associated with building populations. In the first, or pre-epidemic, phase individual trees or small clusters are attacked. In the second, epidemic, phase the clustering attack patterns increase to form large groups of JPB-attacked trees. Coordinates and JPB-attack status were obtained for each tree in a 24.3 ha study area and geospatially rendered to depict the degree of clustering that occurred during each outbreak phase. Tree-level and forest attribute covariates were calculated to represent: tree diameter at 1.4 m in height, distance to nearest beetle source-tree (attacked prior year), beetle source-tree infestation severity in 0.02-ha neighborhood (beetle pressure), and forest density stand density index surrounding each response tree (Egan et al., 2016). Here we focus on the analysis of the JPB clusters formed during the 1993 peak epidemic year and discuss novel approaches to modeling the JPB outbreak dynamics.

Fig. 1 displays live and JPB-attacked trees in the Lake Tahoe Basin area with *the x*- and *y*-axes expressed in the Universal Transverse Mercator (UTM) units. The UTM uses a 2-dimensional Cartesian coordinate system to give locations on the surface of the Earth. To better understand the scale of the plots, the *x*- and *y*-axes in the left panel represent about 500 and 700 m, respectively. In the right panel the *x*- and *y*-axes represent about 40 and 60 m, respectively. The green and red points correspond to the live and JPB-attacked trees, respectively. Recent publications (Egan et al., 2016; Sartwell and Stevens, 1975) both indicate that clusters with more than 5 attacked trees represent spatial clustering typical of epidemic phase of outbreak. Thus, the total of 1139 JPB-attacked trees, were combined into biologically defined groups of $\geq$ 5 spatially aggregated trees using the geospatial ArcGIS software. These groups represented 97% of all trees attacked in 1993. Each cluster was delineated spatially following methods described in Egan et al. (2016) and had its geometric center identified; the cluster centers are denoted by cross marks in Fig. 1. The JPB-attacked groups in the Basin area are marked by red dots in the left panel allowing to visualize the relative size and spatial location of each group. The right panel in Fig. 1 provides a close-up view of Cluster 30.

We are interested in identifying the spatial patterns of JPB-attack and cluster formation within a given year and across multiple years. This could potentially be used to predict which JPB-attack clusters are likely to expand and grow through time. This could potentially be used to predict direction and magnitude of cluster growth based on local forest tree attributes. Given the complex cluster area shapes shown in Fig. 1, it is important to: (1) quantify the size and shape of the affected clusters; and (2) model the association between the shape and size of clusters and various forest attributes.

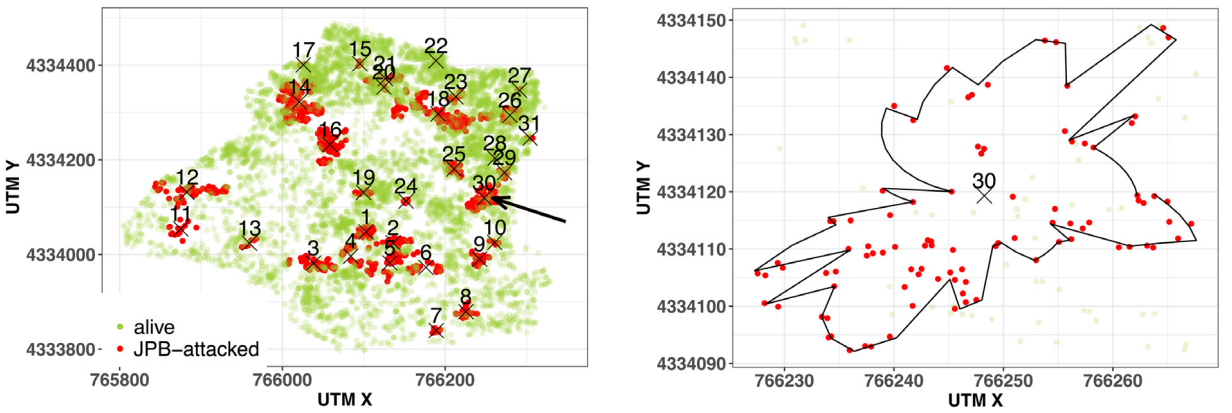

Fig. 1. Left panel: Lake Tahoe Basin Area, dead (red) and alive (green) trees. Each cluster is assigned a numerical label. Cluster centers are indicated with a cross mark. Right panel: zoom into Cluster 30, with the cluster border outlined by the black line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Methods

We now introduce methods for describing the size and shape of the $I = 31$ tree clusters data shown in the left panel of Fig. 1. We have the location of trees affected by JPB, their cluster assignment and the center of each cluster. Based on these data, our objectives were to: (1) derive the contour of each cluster; (2) quantify the contours; and (3) estimate the association between cluster and contour characteristics. To obtain the contour of one cluster we first consider the standard system of coordinates with the origin at the center of the cluster. For each tree we obtain its distance from the cluster center and the angle with respect to the horizontal center line. We separate the coordinate system into non-overlapping 0.1 rad (approximately 5.73°) equal subcomponents. For each circular sector corresponding to an angle of 0.1 rad and the vertex at the center of the cluster, we estimate the cluster border as the maximum distance from the center of all trees within that particular circular sector. Fig. 2 shows the process described above. First, we plot the cluster center $\boldsymbol{c}_i = (c_{xi}, c_{yi})$ and all JPB attacked trees $\boldsymbol{t}_{ik} = (t_{xik}, t_{yik})$, $k = 1, \ldots, K_i$ in the original UTM coordinates. Here i is the cluster number and K_i is the total number of JPB attacked trees in the cluster i. Next, we define the cluster-specific coordinate system with group center $\boldsymbol{c}_i$ at the origin and the kth tree coordinate in cluster i given by $\boldsymbol{t}^c_{ik} = (t^c_{xik}, t^c_{yik}) = (t_{xik} - c_{xi}, t_{yik} - c_{yi})$. We define the distance to the center of the cluster, d_{ik}, and the angle relative to the reference vector $\boldsymbol{v} = (1, 0)$, θ_{ik}, as:

$$d_{ik} = \|\boldsymbol{t}^c_{ik}\|_2 = \sqrt{(t^c_{xik})^2 + (t^c_{yik})^2} \quad \text{and} \quad \theta_{ik} = \cos^{-1}\left(\boldsymbol{v}^t \frac{\boldsymbol{t}^c_{ik}}{\|\boldsymbol{t}^c_{ik}\|_2}\right).$$

The next step is to consider a grid $g = 1, \ldots, 63$ of non-overlapping angles corresponding to 0.1 radians. Let $R(g)$ be the cone centered at the origin and corresponding to the particular angle interval defined by g. If $R(g)$ contains any trees then we define the angle, θ_g, as θ_{ik} rounded to the nearest 1st decimal point and border of the cluster, $d(g)$, as $d(g) = \max_{k \in R_g} d_{ik}$. If there are no trees in R_g then both θ_g and $d(g)$ are missing. Here n_{ig} is the number of affected trees in the sector g. Finally, we linearly interpolate the values of $d_i(\theta_g)$ to observe $d_i(\theta)$ on the equidistant grid $\theta_g = 0, 0.1, 0.2, \ldots, 6.2$ rad (blue dashed line in Fig. 2 step 4).

In addition to the functional representation of the cluster shape, we also have the number $n_i(\theta_g)$ of JPB-attacked trees in the ith cluster at each angle θ_g on the interpolated grid. The number $n_i(\theta_g)$ is strictly positive if there is at least one JPB-attacked tree in this direction, and 0 otherwise. Thus, for each cluster we have derived two functions that have the sector location as argument: the distance to the border of the cluster and the number of affected trees. Visual inspection of these functions for Cluster 30 in Fig. 3 reveals that the cluster is larger in the North-East, $\theta \in (0, \frac{\pi}{2})$, and South-West, $\theta \in (\pi, \frac{3\pi}{2})$, directions. However, the number of affected trees in the North-East is smaller than in the South-West direction. This is just the analysis of one cluster and we would like to investigate whether there are consistent patterns across clusters for each function and whether there is an association between the distance and tree number functions across clusters. We address the first aim by conducting a functional Principal Components Analysis (fPCA) on the smoothed cluster shape and number of trees functions. Viewing the distance $d(\theta)$ and the number of trees $n(\theta)$ cluster functions as $I \times K$ dimensional matrices D and N, where $I = 31$ clusters and $K = 63$ angles $\theta_g = 0, 0.1, \ldots, 6.2$, we can perform fPCA on each data matrix using the fast sandwich smoother (Xiao et al., 2016) **fpca.face** function implemented in **R** package **refund** (Crainiceanu et al., 2013). Results of this analysis are presented in Section 4.

To address the second problem, that is to investigate whether there is an association between the distance and tree number functions across clusters, we focus on function on function regression, where the exposure is the number of trees and the outcome is the distance to the boundary of the cluster. We modeled, at every angle θ, the cluster shape $d_i(\theta)$ as a

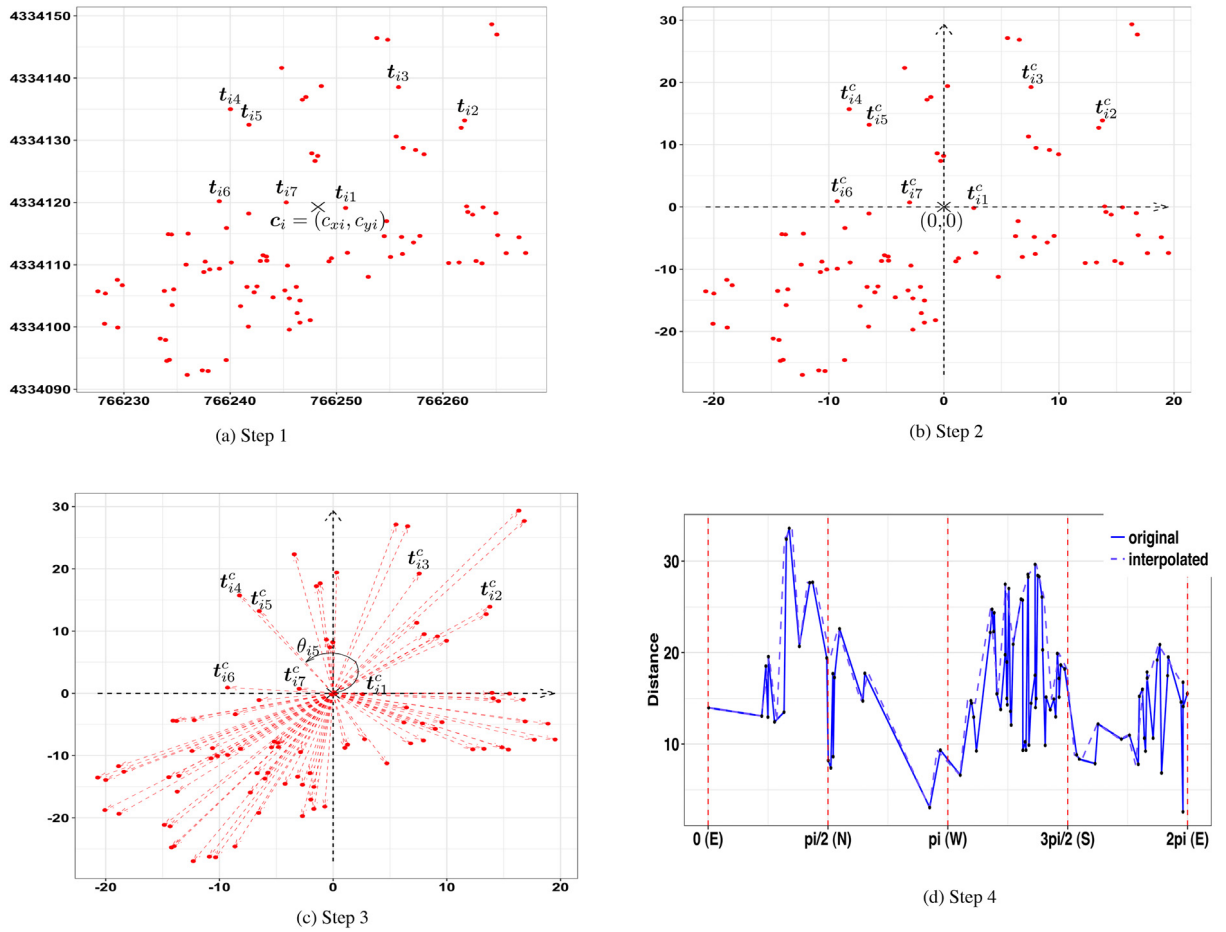

Fig. 2. Step-by-step illustration of the ith tree cluster representation as a function $d_i(\theta)$.

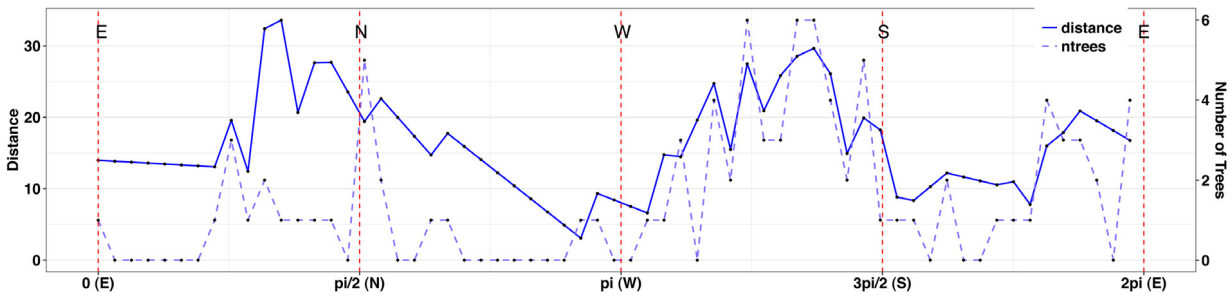

Fig. 3. Distance from cluster center $d(\theta)$ (solid line) and the number of trees $n(\theta)$ (dashed line) functions for cluster number 30.

function of the number of trees $n_i(\cdot)$, and forest attributes covariates x_{ik}, $k = 1, \ldots, p$ that do not change with θ. This can be done using the function-on-function regression model with log-link

$$\log(\mathbb{E}\{d_i(\theta)\}) = \mu(\theta) + \int_{s=0}^{s=2\pi} \{n_i(s) - \bar{n}_i\}\beta(\theta, s)ds + \sum_{k=1}^{p} x_{ik}\beta_k, \tag{2}$$

where $\mu(\theta)$ is the overall intercept, $\bar{n}_i$ is the ith cluster average number of trees, $\beta(\theta, s)$ is the unknown regression surface weight placed on the number of trees at angles $\theta_g = 0, 0.1, \ldots, 6.2$ rad to predict cluster distance at angle θn $\beta_k, k = 1, \ldots, p$, are the effects of forest attributes, and $\epsilon_i(\theta)$ are t-distributed error terms. Model (2) is the special case of the

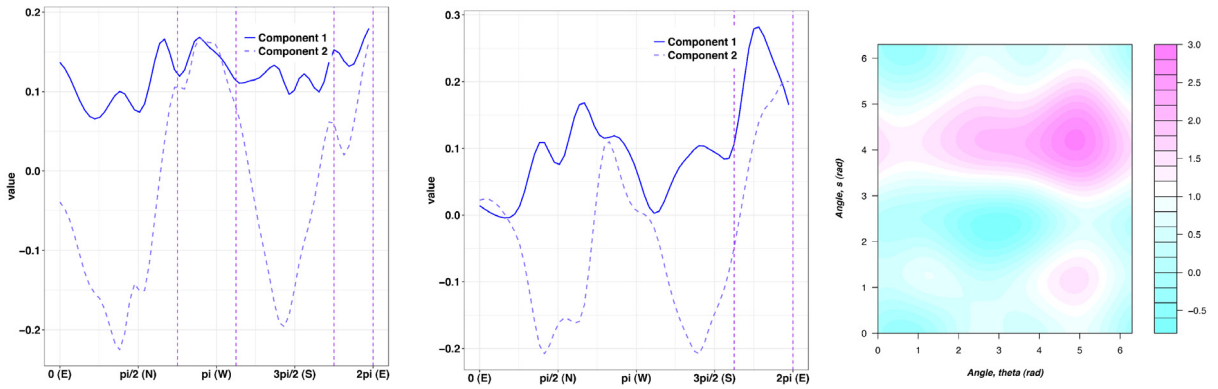

Fig. 4. First and second principal components for the cluster shape (left panel) and the number of trees (middle panel) covariance functions. Vertical dashed lines indicate directions of maximum variability. Right panel: the estimated regression surface $\beta(\theta, s)$.

Functional Additive Mixed Models (Scheipl et al., 2015) and can be fitted via penalized function-on-function regression **pffr** function implemented in the **R** package **refund** (Crainiceanu et al., 2013).

4. Results

The first 2 principal components for the distance and number of trees functions are presented in the left and right panels of Fig. 3, respectively. For the distance functions the first two components explain approximately 82% (Component 1: 65% and Component 2: 17%). For the number of trees functions the first two components explain 80% (Component 1: 56% and Component 2: 24%) of the total variability. The first component of the distance function is positive and relatively close to a constant, indicating that this is a surrogate for cluster size. The second component indicates a contrast between the North and South directions on one side and the West and East directions on the other side. Clusters that are positively loaded on this PC will tend to be much more elongated in the West–East direction than in the North–South direction. Clusters that are negatively loaded on this component will tend to be more elongated in the North–South direction. The first principal component of the number of trees affected is also positive indicating that the difference between clusters is mainly in terms of total number of trees affected. However, the first PC is not flat, indicating that the North direction and South-East directions are the ones where most differences between the number of affected trees will be accumulated. The second principal component for the number of trees is almost identical to the second principal component for the size of the cluster.

We employed the function-on-function regression model (2) with the distance function as outcome and number of trees as functional predictor. We considered four cluster-level covariates obtained by averaging within cluster the following tree-specific characteristics: average tree diameter (x_{i1}), distance to nearest beetle source tree (x_{i2}), beetle pressure in a 0.02-ha neighborhood (x_{i3}), and forest density (x_{i4}). The proportion of deviance explained by the functional regression was 41.4% and both the cluster-specific effects and the functional effect were highly significant (p-value < 0.001). The χ^2-scores and corresponding estimated degrees of freedom for the 4 cluster-level covariates were: 919.22 (7 df), 52.12 (5 df), 1027.46 (8 df), 87.19 (8 df), respectively. Fig. 4 right panel shows the heatmap of the estimated mean regression surface $\beta(\theta, s)$ of the number of trees along angles θ. For example, consider the angle $\theta = 4.71$ (South). The weight function, $\beta(\theta, s)$, is strongly positive in the West to South to East (s=3.14 to 5.5) indicating that, for a fixed average density of trees in the cluster, higher density of trees in the West, South, and South-East directions had a strong effect on increasing the cluster size in the South direction.

5. Discussion

By expressing the complex JPB-attacked cluster regions as functions of the direction from the cluster center, we have developed a method for modeling the association between the shape and size of clusters and various forest attributes. This approach allows to use functional data modeling to quantify the directions of beetle expansion and predict cluster size as a function of spatial direction. We used functional principal component analysis to examine the variability of the cluster-induced functional data and detect patterns across cluster sizes and shapes. The first and second functional principal components were surrogates for cluster size and the contrast between geographic directions of cluster expansions, respectively. Major differences in the number of affected trees were found to be in the North and South-East directions, which may indicate the preferred geographical directions of beetle attack. We have also quantified the association between the distance and tree number functions across clusters using function-on-function regression. In summary, the methods presented in this paper provide a class of modeling tools for quantifying the preliminary observations discussed in Egan et al. (2016).

One limitation of the approach is that the beetle expansion in direction θ is limited to the year 1993. This does not provide information across years and it is not currently designed to help with prediction for the next years. Generalizing our approach to take into account temporal dynamics during the 1991–1996 is a natural next methodological step. Another major methodological area would be to construct predictive models and incorporating global directions of beetle-attack and determine the potential affect of cluster location, shape and size on beetle expansion.

Finally, while the methodology discussed in this paper is motivated by the analysis of clusters formed by JPB attacked trees, it is not limited to this application. The proposed analysis can be used in a number of ecological and health applications, including those described in Duczmal et al. (2006), Yiannakoulias et al. (2010), Duarte et al. (2010), Moreira et al. (2015) and Lawson (2010), after the data clusters have been outlined using either domain expert knowledge or existing cluster detection algorithms.

References

Bhattacharjee, A., Cai, L., Maiti, T., 2017. Functional regression over irregular domains: variation in the shadow price of living space. Spat. Econ. Anal. 12 (2–3), 182–201.

Crainiceanu, C.M., Reiss, P.T., Goldsmith, J.A., Huang, L., Huo, L., Scheipl, F., 2013. refund: regression with functional data.

Duarte, A.R., Duczmal, L., Ferreira, S.J., Cançado, A.L.F., 2010. Internal cohesion and geometric shape of spatial clusters. Environ. Ecol. Stat. 17 (2), 203–229.

Duczmal, L., Kulldorff, M., Huang, L., 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. J. Comput. Graph. Statist. 15 (2), 428–442.

Egan, J.M., Sloughter, J.M., Cardoso, T., Trainor, P., Wu, K., Safford, H., Fournier, D., 2016. Multi-temporal ecological analysis of Jeffrey pine beetle outbreak dynamics within the Lake Tahoe Basin. Popul. Ecol. 58, 441–462.

Kerr, W.R., Kominers, S.D., 2015. Agglomerative forces and cluster shapes. Rev. Econ. Stat. 97 (4), 877–899.

Lawson, A.B., 2010. Hotspot detection and clustering: ways and means. Environ. Ecol. Stat. 17 (2), 231–245.

Marcon, E., Puech, F., 2017. A typology of distance-based measures of spatial concentration. Reg. Sci. Urban Econ. 62, 56–67.

Moreira, G.J., Paquete, L., Duczmal, L.H., Menotti, D., Takahashi, R.H., 2015. Multi-objective dynaic prograing for spatial cluster detection. Environ. Ecol. Stat. 22 (2), 369–391.

Mori, T., Smith, T.E., 2013. A probabilistic modeling approach to the detection of industrial agglomerations. J. Econ. Geogr. 14 (3), 547–588.

Ramsay, T., 2002. Spline smoothing over difficult regions. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (2), 307–319.

Sangalli, L.M., Ramsay, J.O., Ramsay, T.O., 2013. Spatial spline regression models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 681–703.

Sartwell, C., Stevens, R.E., 1975. Mountain pine beetle in ponderosa pineprospects for silvicultural control in second-growth stands. J. Forestry 73 (3), 136–140.

Scheipl, F., Staicu, A.-M., Greven, S., 2015. Functional additive mixed models. J. Comput. Graph. Statist. 24 (2), 477–501.

Xiao, L., Ruppert, D., Zipunnikov, V., Crainiceanu, C., 2016. Fast covariance estimation for high-dimensional functional data. Stat. Comput. 26, 409–421.

Yiannakoulias, N., Wilson, S., Kariuki, H.C., Mwatha, J.K., Ouma, J.H., Muchiri, E., Kimani, G., Vennervald, B.J., Dunne, D.W., 2010. Locating irregularly shaped clusters of infection intensity. Geospatial Health 4 (2), 191–200.